Machine Learning Literature Review

Machine Learning in Stock Markets

UCL Connect.ed Mentorship Project 2021

Lau Pak To Ryan
15/2/2021

Abstract

Something something kekw

Contents

1 Introduction

Stock markets are known to be disordered and highly unpredictable. The up and down movements of the market depends on many different factors. However, from individuals who want to earn money out of stock exchange, to the big companies who adjust their business and trading strategies on the movements of the stock, it is extremely clear that the prediction of stock prices are of great interest. In this literature review, we will be studying some of the methodologies, findings and limitations of using ML techniques to predict the movements of the stock market as a preliminary task. With reference to the reviewed literature, we aim to know which ML techniques are used, what are their advantages and disadvantages, where to collect data, and areas which are not explored. Ultimately, we hope that this review will give us an insight of the potential topics or questions we might be able to answer using our own methodologies and models in the future.

2 Review of literature

4 main pieces of literature were reviewed. In the following section, we will discuss each of their methodologies, ML techniques and findings. Limitations and possible future work proposed by the authors will be included as well. Each paper are simply catergorised with the ML technique used in the study.

2.1 Adaptive Boosting (Adaboost)

2.1.1 General Review

In the first paper of which we are discussing, P. Rodriguez and S. Sosvilla River [1] try to find a pattern which shows the movements and predictability of the stock market using ML techniques. The particular algorithm they used is called adaptive boosting, or Adaboost. It aims to improve learning performance and was developed by Yoav Freund and Robert Shapire in 1995 [2]. The researchers used data from S&P daily closing prices and divided the data set into 4 non-overlappingsets, and further divided each set into two sub-samples. The former is used for training and learning, and the latter is used for testing. It was assumed that the movements of the future stock price were associated to past returns. The results suggested that Adaboost was able to find a function which differentiated the ups and downs of the stock market, better than randomly made decisions.

2.1.2 Limitations and Future Work

Adaboost fails to measure the predictability of sample which are outside the learning data. The authors noted that ML algorithms can be altered to study large proce movements, which are helpful for choosing trading strategies, of which we will see a study [5] investigating this in a future section. Moreover, ML algorithms can be utilized to look for risk exposures.

2.2 Boosted Regression Trees (BRT)

2.2.1 General Review

The next paper discusses an ML algorithm known as boosted regression trees (BRT) to run tests in order to check whether we can make predictions on stock markets [3]. This algorithm selects the predictor variables itself to measure forecast errors more accurately while accounting for implicit non-linear dependence or other dependencies between forecast errors and predictor variables. Similar methods have been used in a study conducted by Döpke, Fritsche, and Pierdzioch to predict economic recessions [4]. The orthogonality test was used to model forecast errors. Individual trees were estimated iteratively. A strong learner was then added to a group of simple base learners generated by the iterative process. The resultant trees were then used to model forecast errors. The researchers of the paper aimed to "re-examine the orthogonality property of forecast errors implied by aggregate stock market forecasts as represented by a consensus forecast implied by the stock market forecasts from the Livingston Survey" [3]. If it is possible to foretell the forecast error, then the rational expectation hypotheses REH) should be rejected. Using BRT, results have shown that the REH could be accepted for short term forecasts, and the method proposed was useful for modelling the study of forecast errors.

2.2.2 Limitations and Future Work

It is suggested by the researchers that study should be carried out whether well-known predictor variables can predict the ups and downs of the forecast error implied by the stock market forecasts, and it may be attainable to exploit these predictive values and come up with respective trading strategies in the stock market. On the other hand, the authors recommended that future studies could be conducted using the same BRT algorithm used, however using different regression or classification trees.

2.3 Logistic Regression (LR), Random Forests (RF) and Support Vector Machine (SVM)

2.3.1 General Review

A study conducted in Korea also used ML to attempt predicting the global stock market. Building on these estimates, they suggested practical strategies for investments in the global market [5] . They aim to examine the effectiveness and improvement of their strategies on a financial network based on global stock indices of 10 countries. These 10 countries are selected using the Morgan Stanley Capital International (MSCI) Index countries and the regional representation. Five of the them were categorised as developed market (DM) and the other 5 countries were categorised as emerging market (EM), with accordance to the MSCI world index market classification. The performance of the proposed approach was evaluated based on global investment strategy, which consists of different stock volatility and prediction periods. Volatility is defined as "a statistical measure of the dispersion of returns for a given security or market index." [6] Volatility increases with risk in security. They also incorporated the structural changes in financial networks by comparing a stable period in 1996, and 2 turnmoil periods, or disordered periods, each in 1997 and 2008. 3 ML techniques were used in this research. Logistic Regression (LR) is a forecasting model used to guess the movements if stocks and stock price indices. It had been said to be useful when the measured variable is binary or multinomial, and has been implemented in the field of investments often. The other technique used was Random Forests (RF), which is an extension of traditional decision tree techniques. It is made up of randomly selected variables, which are an addition of multiple decision trees, which were expected to increase the accuracy and stability of the constructed model. Over-fitting was avoided by increasing the size of the forest, as this can generalize the error to a fixed value. This technique has been proven to be the best in predicting imbalanced data based on the study conducted by Brown and Mues [7]. The last technique was the Support Vector Machine (SVM), which are well suited for classification problems, as it minimizes over-fitting, is high in accuracy, easy to modify and able to grasp high-dimensional data [8]. The results showed that the global market indicators are effective and useful for market forecasting during the market turnmoil period. The best strategy was to make mid-term investments with short term volatility. The short term volatility had been shown to be better amongst the other volatility periods as short-term is more sensitive to change.

2.3.2 Limitations and Future Work

It is shown that the SVM learning process for determining the best classifier in a given data consumes too much time, due to the fact that it contains a large computational complexity. The study has minimal studies on practical usefulness. Different to investment simulations, averages do not suffice practically. It is crucial to improve the techniques for selecting the parameters to obtain more practical simulation results in future works. The researchers have not used other network indicators other than connectedness measure to capture varying dynamics of the stock market. 3 ML techniques were not sufficient in measuring the effects of various implicit factors in markets, which are very complex. The authors suggested to use model improvement or deep learning in future studies to tackle this problem. Other important factors such as economic indicators have not been used.

2.4 Hybrid Method of Genetic Algorithm (GA) and Artificial Neural Network (ANN)

2.4.1 General Review

A study performed by O. M. E. Ebadati and M. T. Mortazav [9] is quite different from the literature described previously. They administered two types of ML algorithm together, which is a hybrid method to predict stock prices. The researchers aimed to develop a method predicting stock prices and time series. Similar to the study conducted by P. Rodriguez and S. Sosvilla River [1], they assumed also in this paper that the future stock price can be predicted using previous stock prices. The model was based on the opening and closing of the stock prices. The two techniques are so called Genetic Algorithm (GA) and Artificial Neural Network (ANN). The GA finds and initialises the weights of the ANN algorithm, and ANN sets errors on a fixed number with fewer computational repeats to reduce time consumption. Back propagation (BP) was implemented to train ANN to find the desired pattern. The method applied was proven to be highly successful as high accuracy was reached in less time. By applying the method to the Apple stocks dataset, it reached a 99.99% improvement in minimizing the error function and 90.66% improvement in time, along with other promising case studies.

2.4.2 Limitations and Future Work

It is recommended by the researchers that a hybrid method can be used to tackle more finance problems. For example, using the proposed hybrid method with SVM as used in the previous paper [5] .

3 Potential research topics and methodologies

As seen from the literature, it is perfectly feasible to predict stock market trends using past data. As for our research topic, data must be collected from a reliable source. For instance, the Thompson Reuters DataStream Database as used by T. Lee, J. Cho, D. Kwon and S. Sohn [5]. The methods can potentially be used individually [5] or hybrid [9]. We can also compare which technique will be the best in forecasting trends of the stock market. We could also find the correlations between the variables affecting forecast errors, which are discussed by C. Pierdzioch and M. Risse [3]. Some general research topics can be extensions or modifications of the existing literature, with different sets of data and assumptions. Comparisons with similar literatures can be made. A more practical tool we can come up with is to plan when to buy or sell a particular stock for companies and individuals of the general public, which tackles the problem of practicality as discussed [5] . Some recent topics regarding the stock market can also be potential topics. For example, the recent Gamestop stock [10], where we could ask the question whether bubble bursts could be predicted using ML techniques. The exact focus and methodologies should be refined in future group discussions.

4 Conclusion

In conclusion, as seen from the reviewed studies, it is highly feasible and effective to implement ML techniques and algorithms to forecast or predict stock market behaviours. We could use the limitations as described to decide and refine on a specific research question. The specifics of the research topic can be discussed in the near future.

On a final note, it is important to recognise that data simulated from any ML techniques possibly do not reflect the real life stock exchange market, due to its complexity.

References

[1] P. Rodriguez and S. Sosvilla Rivero, "Using Machine Learning Algorithms to Find Patterns in Stock Prices", SSRN Electronic Journal, 2006. Available: 10.2139/ssrn.893141

[2] Y.Freund, R.Shapire. A decision-theoretic generalization of on-line learning and application to boosting. Proceedings of the Second European Conference on Computaional Learning Theory, 1995:23-27

[3] C. Pierdzioch and M. Risse, "A machine-learning analysis of the rationality of aggregate stock market forecasts", International Journal of Finance Economics, vol. 23, no. 4, pp. 642-654, 2018. Available: 10.1002/ijfe.1641

[4] Döpke, J., Fritsche, U., Pierdzioch, C. (2017). Predicting recessions in Germany with boosted regression trees. International Journal of Forecasting, 33, 745–759

[5] T. Lee, J. Cho, D. Kwon and S. Sohn, "Global stock market investment strategies based on financial network indicators using machine learning techniques", Expert Systems with Applications, vol. 117, pp. 228-242, 2019. Available: 10.1016/j.eswa.2018.09.005

[6] [1]"Volatility", Investopedia, 2021. [Online]. Available:https://www.investopedia.com/terms/v/volatility.asp#:~:text=Volatility%20is%20a%20statistical%20measure,same%20security%20or%20market%20index.

[7] G. Brown and M. Cliff, "Investor Sentiment and Asset Valuation", The Journal of Business, vol. 78, no. 2, pp. 405-440, 2005. Available: 10.1086/427633

[8] [C. Cortes and V. Vapnik, "Support-vector networks", Machine Learning, vol. 20, no. 3, pp. 273-297, 1995. Available: 10.1007/bf00994018

[9] O. M. E. Ebadati and M. T. Mortazavi, "AN EFFICIENT HYBRID MACHINE LEARNING METHOD FOR TIME SERIES STOCK MARKET FORECASTING: INTERNATIONAL JOURNAL ON NEURAL AND MASS - PARALLEL COMPUTING AND INFORMATION SYSTEMS," Neural Network World, vol. 28, (1), pp. 41-55, 2018. Available: https://www-proquest-com.libproxy.ucl.ac.uk/scholarly-journals/efficient-hybrid-machine-learning-method-time/docview/2012893389/se-2?accountid=14511. DOI: http://dx.doi.org.libproxy.ucl.ac.uk/10.14311/NNW.2018.28.003.

[10] https://finance.yahoo.com/news/gamestop-continues-fall-short-squeeze-135358733.html